Análise de sobrevivência

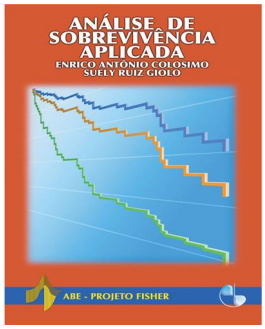
Professor: Pedro Monteiro de Almeida Junior

Conteúdo programático

1. **Conceitos básicos em análise de dados de sobrevivência**
   - Tempos de falha e censura
   - Exemplos de uso da análise de sobrevivência
   - Representação de dados de sobrevivência
   - Funções de interesse e suas relações
2. **Técnicas não paramétricas para análise de sobrevivência**
   - Estimador não paramétrico de Kaplan-Meier
   - Comparação entre curvas de sobrevivência
3. **Métodos paramétricos para análise de sobrevivência**
   - Distribuições básicas em sobrevivência (Exponencial, Weibull e lognormal)
   - Procedimentos para a escolha de uma distribuição
   - Exemplos práticos no R
4. **Modelos de regressão para a análise de dados de sobrevivência**
   - Modelo de regressão exponencial
   - Modelo de regressão Weibull
   - Modelo de regressão Lognormal
   - Análise de Resíduos
   - Exemplos práticos no R
5. **Modelo de riscos proporcionais de Cox**
   - Modelo de Cox
   - Estimação e interpretação dos parâmetros
   - Análise de resíduos
   - Exemplos práticos no R

Referência bibliográfica principal

[1] E. A. Colosimo, S. R. Giolo, ANÁLISE DE SOBREVIVÊNCIA APLICADA , Edgard Blucher, São Paulo, 2006.



Introdução a análise de sobrevivência

O pontapé inicial da análise de sobrevivência foi na área da saúde. Entretanto, ela se aplica em diversas outras áreas de aplicação. É comum se deparar com situações que é necessário obter o tempo até a ocorrência de determinado evento. Por exemplo:

- Tempo até a cura de um paciente
- Tempo até a inadimplência de um cliente
- Tempo até a ocorrência de fraude
- etc ...

Dessa forma, conseguimos determinar a sobrevida de um paciente, produto, clientes, etc... Aqui no curso, o objetivo é avaliar as principais técnicas de análise de sobrevivência e aplicá-las em diversas áres do nosso conhecimento. A sequir, vamos ver os principais conceitos em sobrevivência.

CONCEITOS INICIAIS

Em análise de sobrevivência, a nossa variável de interesse é o tempo até a ocorrência de um determinado evento. Esse tempo é o que chamamos de **tempo de falha**. Na área da saúde por exemplo, o tempo de falha poderia ser: o tempo até a cura de um paciente, tempo até a reincidência de uma doença ou tempo até o falecimento do paciente.

Os dados para análise de sobrevivência envolvem uma resposta que é avaliada sobre um período de tempo, que na maioria das vezes será difícil acompanhar todas as observações até a ocorrência do evento. Dessa forma teremos no estudo a presença de observações incompletas. Daí vem o conceito de censura nos dados.

Censura: a principal característica de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta. Por exemplo, em algumas situações, o acompanhamento do indivíduo presente no estudo é interrompido (o indivíduo mudou de cidade, faleceu, o estudo finalizou antes de ocorrer o evento desejado, etc. . .).

Observação: Sem a presença de censura, as técnicas estatísticas clássicas, como análise de regressão e análise de experimento poderiam ser usadas na análise desse tipo de dado. Por exemplo:

Suponha que o interesse seja o de comparar o tempo médio de vida de três grupos de pacientes. Se não houver censuras, podemos usar a análise de variância para fazer a comparação entre os grupos de pacientes usando o tempo médio de vida. Entretanto, se houver censuras, não teremos a informações de pacientes que não concluíram o estudo. Portanto o uso de técnicas de análise de sobrevivência se torna indispensável nesses casos.

Caracterizando dados de sobrevivência

Os dados de sobrevivência são caracterizados por: (i) tempos de falha e (ii) censuras. Estas duas características constituem a nossa variável resposta (ou de interesse). A seguir vamos discutir essas características individualmente.

Tempo de Falha

Tempo de Sobrevivência ou Falha: Tempo até a ocorrência do evento de interesse.

O tempo de falha é constituído por três elementos:

Elementos que definem o tempo de falha:

1. A escala de medida
2. O tempo inicial
3. O evento de interesse

O tempo de sobrevivência vai do tempo inicial até a ocorrência do evento de interesse usando a escala de medida definida.

Censura

É importante ressaltar que mesmo sendo incompletas as observações censuradas nos fornecem informações sobre os tempos de falha. A omissão dessas observações no cálculo das estatísticas de interesse pode acarretar em conclusões viciadas.

TIPOS DE CENSURA:

1. **Censura à direita:** O tempo de falha é superior ao tempo registrado. Desprezar essa informação faria com que o risco de ocorrência do evento de interesse fosse superestimado pois o tempo até a falha é desconhecido mas o evento de interesse não ocorreu até o último momento observado. Existem três conhecidos mecanismos de censura à direita.
2. **Censura do Tipo I:** O estudo será terminado após um período préestabelecido de tempo. As observações cujo evento de interesse não foi observado até este tempo são ditas censuradas.
3. **Censura do Tipo II:** O estudo será terminado após ter ocorrido o evento de interesse para um número pré-estabelecido de observações.
4. **Censura Aleatória:** Ocorre se a observação for retirada no decorrer do estudo sem ter ocorrido o evento de interesse ou se o evento de interesse ocorrer por uma razão diferente da estudada.
5. **Censura Intervalar:** ocorre quando o evento ocorreu em um certo intervalo. Isto é, o tempo exato da falha não é conhecido exatamente mas pertence a um intervalo. Ocorre por exemplo em estudos em que pacientes são acompanhados em visitas periódicas.

   Ilustração dos tipos de censura:

Representação dos dados de sobrevivência

Os dados de sobrevivência para o indivíduo i ($i = 1, \ldots, n$) sob estudo são representados, em geral, pelo par (T_i, C_i), sendo T_i o tempo de falha ou censura e C_i a variável indicadora de falha ou censura, isto é,

$$C_i = \begin{cases} 1, & \text{se } T_i \text{ é um tempo de falha} \\ 0, & \text{se } T_i \text{ é um tempo censurado} \end{cases}$$

Assim, a variável de interesse em análise de sobrevivência é representada por duas colunas (Tempo até a falha e censura).

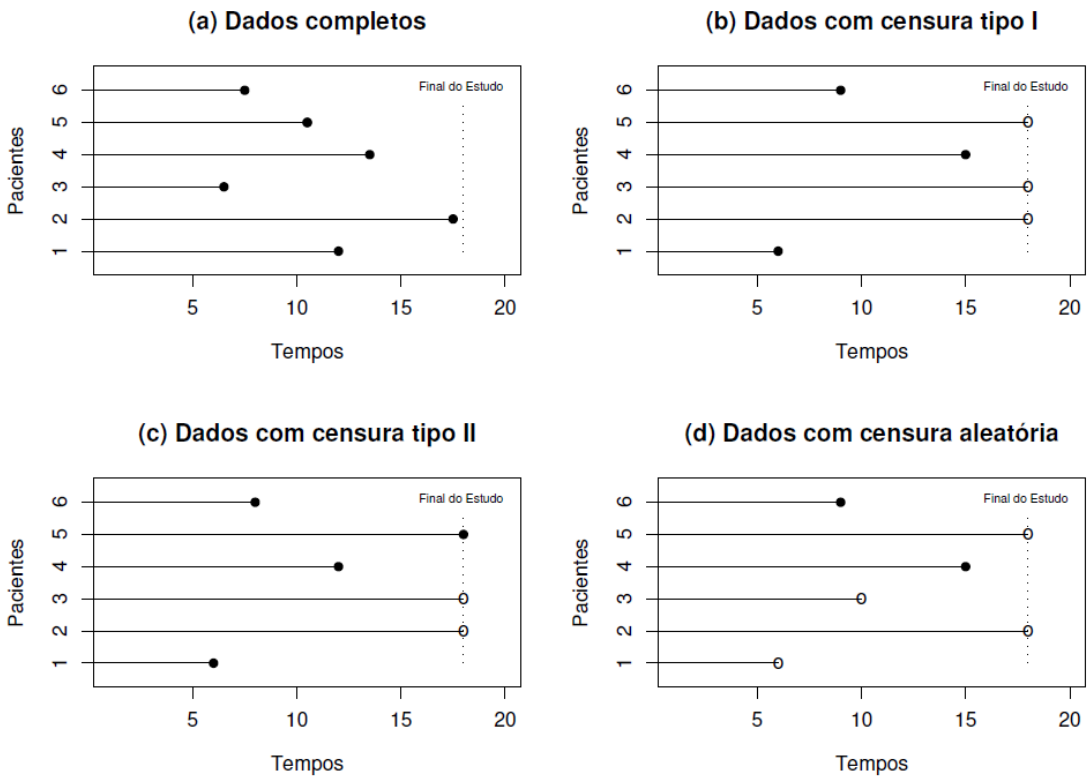


Figure 1: Fonte: Colosimo & Giolo (2006)

Estrutura dos banco de dados de sobrevivência

Indivíduo (i)	Tempo (T_i)	Censura (C_i)	covariável (X_i)
1	10	1	5
2	15	0	6
3	30	1	8

Exemplos de Dados de Sobrevivência

Em diversas situações podemos usar técnicas de análise de sobrevivência. A área de maior destaque para essa técnica é a área médica. Podemos usar a análise de sobrevivência para identificar fatores de prognóstico para uma doença, comparação de tratamentos (uma prática bastante comum na área de oncologia, no qual podemos testar se um novo tratamento é melhor que o tratamento usual, ou seja, podemos avaliar se o tempo de sobrevivência de pacientes que aderiram ao novo tratamento é maior). A seguir vamos apresentar alguns exemplos da aplicação de sobrevivência.

1. Dados de Hepatite

Um estudo clínico aleatorizado foi realizado para investigar o efeito da terapia com esteróide no tratamento de hepatite viral aguda. 29 pacientes com esta doença foram aleatorizados para receber um placebo ou o tratamento com esteróide. Cada paciente foi acompanhado por 16 semanas ou até a morte (evento de interesse) ou até a perda de acompanhamento. Os tempos de sobrevivência observados, em semanas, para os dois grupos são apresentados são dados a seguir:

Table 2: Tabela: Tempos em dia observados no estudo da hepatite

Grupos	Tempo	Censura
Controle	1	0
Controle	2	0
Controle	3	1
Controle	3	1
Controle	3	0
Controle	5	0
Controle	5	0
Controle	16	1
Controle	16	1
Controle	16	1
Controle	16	1

Grupos	Tempo	Censura
Controle	16	1
Controle	16	1
Controle	16	1
Controle	16	1
Esteróide	1	1
Esteróide	1	1
Esteróide	1	1
Esteróide	1	0
Esteróide	4	0
Esteróide	5	1
Esteróide	7	1
Esteróide	8	1
Esteróide	10	1
Esteróide	10	0
Esteróide	12	0
Esteróide	16	1
Esteróide	16	1
Esteróide	16	1

Outros exemplos:

1. *Em experimentos padrões na investigação de substâncias cancerígenas, animais em laboratórios são sujeitos a doses de substâncias e então são observados para ver se eles desenvolvem tumores. A principal variável de interesse pode ser o tempo até a ocorrência de tumores ou o tempo até a morte.*
2. *Itens manufaturados de componentes eletrônicos são frequentemente sujeitos a testes de vida para obter informações sobre sua duração. Em laboratórios os itens são observados até a falha.*
3. em análise de crédito, podemos estudar o comportamento de clientes para identificar fatores que impactam o tempo até a primeira inadimplência.

Funções usadas para especificar o tempo de sobrevivência

O tempo de sobrevivência é uma variável aleatória T, contínua e positiva. Sendo assim, podemos atribuir alguma distribuição de probabilidade para T que pode ser especificada por algumas funções usadas na análise de sobrevivência, dentre as principais estão a função de sobrevivência e taxa de falha. A seguir, vamos apresentar algumas dessas funções.

1. Função de densidade

No contexto de sobrevivência a função $f(t)$ pode ser interpretada como limite da probabilidade de um indivíduo sofrer um evento em um pequeno intervalo por unidade de tempo.

Esta função é definida por

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t)}{\Delta t}$$

Na **ausência de censuras** a f.d.p. pode ser estimada como a proporção de indivíduos que falham em um intervalo por unidade de tempo.

$$\begin{aligned} \hat{f}(t) &= \frac{\text{Número de ocorrências em } (t, t + \Delta t)}{(\text{Número total de ocorrências }) \times (\Delta t)} \\ &= \frac{N(t)}{(\text{Número total de ocorrências }) \times (\Delta t)} \end{aligned}$$

2. Função de sobrevivência

A **função de sobrevivência** é uma das mais importantes e usadas funções em análise de sobrevivência. A função de sobrevivência é definida como a probabilidade de uma observação não falhar (sobreviver) até um certo tempo t. Em termos probabilísticos, isto é escrito como:

$$S(t) = P(T \geq t).$$

Dessa forma, a distribuição acumulada é definida como:

$$P(T \leq t) = 1 - S(t)$$

ou seja,a **distribuição acumulada** é a probabilidade de uma observação **não sobreviver** ao tempo t. Na Figura abaixo, temos um exemplo de comparação de curvas de sobrevivência entre dois grupos.

Como podemos ver, as curvas de sobrevivência para os dois grupos são diferentes. Em 10 anos, a taxa de sobrevivência para o grupo 1 é de aproximadamente 0.9, enquanto que para o grupo 2 é de 0.5.

Podemos estimar a curva de sobrevivência, quando não há censuras, da seguinte forma:

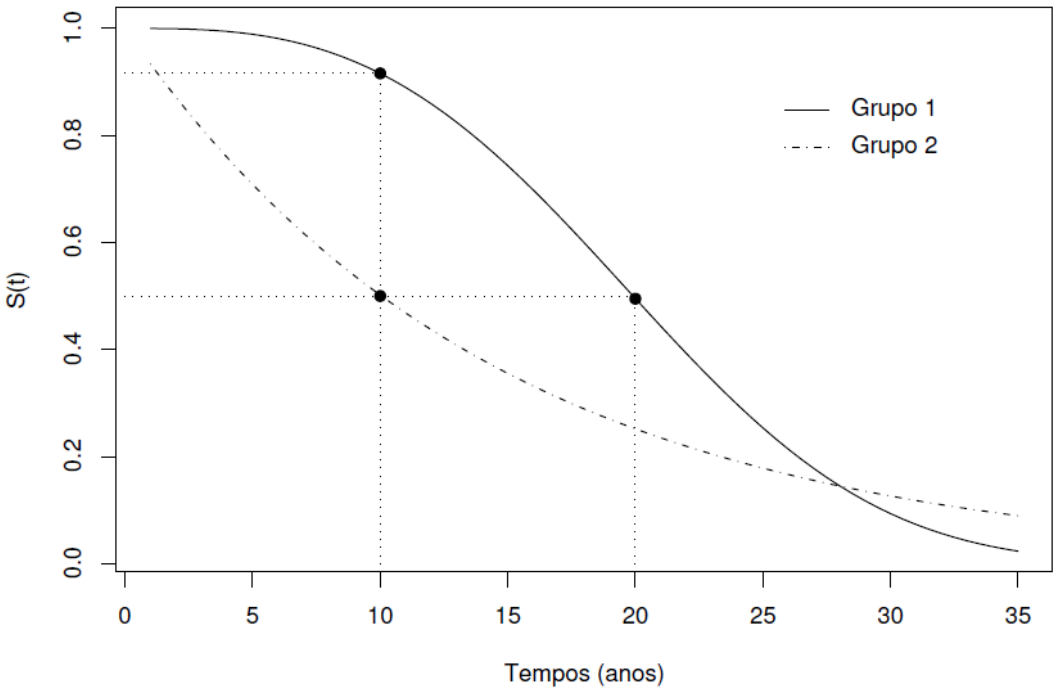


Figure 2: Fonte: Colosimo & Giolo (2006)

$$\hat{S}_x(t) = \frac{\text{Número de indivíduos sob risco no início do intervalo de tempo x}}{\text{Número total de indivíduos}}$$
$$= \frac{R(t)}{\text{Número total de indivíduos}}$$

3. Função de Taxa de Falha (ou risco):

A função taxa de falha é definida como a probabilidade de um indivíduo sofrer o evento entre o tempo t e $t + \Delta t$, dado que ele sobreviveu até o tempo t.

Expressa o risco instantâneo de ocorrência de um evento em um pequeno intervalo de tempo, dado que até então o evento não tenha ocorrido.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

IMPORTANTE: h(t) é uma taxa e não uma probabilidade, pode assumir qualquer valor positivo.

Na ausência de censura a função de risco é estimada por

$$\hat{h}(t) = \frac{N(t)}{R(t) \times \Delta t}$$

N(t) : Número de eventos observados em cada intervalo de tempo (iniciando em t).

$R(t)$: Número de observações sob risco no início do intervalo.

Δt : amplitude do intervalo.

Comportamento da função de risco

A função de risco permite analisar o risco de um indivíduo sofrer um evento em um determinado tempo t, dado que ele já sobreviveu até aquele momento. Por exemplo, qual é o risco de um cliente se tornar inadimplente logo após a liberação de um crédito ? Será que o risco é constante ao longo do tempo ? É provável que não. Dessa forma, podemos ter vários tipos de comportamento de risco.

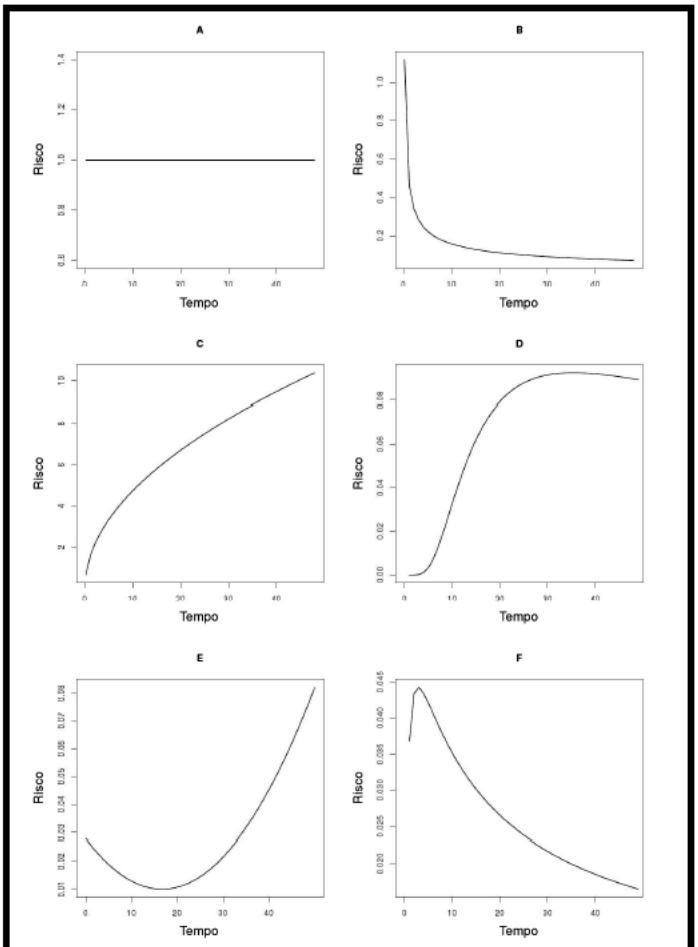


Figura 3.3: Função de risco com diversos formatos

Figure 3: Fonte: Carvalho et al. (2019)

Função de risco (falha) acumulado

A função de risco acumulado $\Lambda(t)$ mede o risco de ocorrênia do evento até um determinado tempo t. Matematicamente, isso significa a soma de todos os riscos em todos os tempos até o tempo t. Como o tempo é uma variável aleatória contínua, T, o risco acumulado é dado por:

$$\Lambda(t) = \int_0^t \lambda(u)d(u)$$

Relações entre as funções

Para T uma variável aleatória contínua e não negativa, tem-se, em termos das funções definidas anteriormente, algumas relações matemáticas são estabelecidas:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}[\log S(t)],$$

$$\Lambda(t) = -\log S(t),$$

$$S(t) = \exp\{-\Lambda(t)\}$$

Portanto, com o conhecimento de uma das funções, por exemplo $S(t)$, implica no conhecimento das demais funções, isto é, $F(t)$, $f(t)$, $\lambda(t)$ e $\Lambda(t)$.

Exercícios

1. Um grande número de indivíduos foi acompanhado para estudar o aparecimento de um certo sintoma. Os indivíduos foram incluídos ao longo do estudo e foi considerado como resposta de interesse a idade em que este sintoma apareceu pela primeira vez. Para os seis indivíduos selecionados e descritos a seguir, identifique o tipo de censura apresentado.

(a) O primeiro indivíduo entrou no estudo com 25 anos já apresentando o sintoma.

(b) Outros dois indivíduos entraram no estudo com 20 e 28 anos e não apresentaram o sintoma até o encerramento do estudo.

(c) Outros dois indivíduos entraram com 35 e 40 anos e apresentaram o sintoma no segundo e no sexto exames respectivamente, após terem entrado no estudo. Os exames foram realizados a cada dois anos.

(d) O último indivíduo selecionado entrou no estudo com 36 anos e mudou da cidade depois de 4 anos sem ter apresentado o sintoma.

Exemplo no R

Carregando as bibliotecas:

```
library(magrittr)
library(dplyr)
```

Criando o dataset sem dados censurados

```
### Criando o dataset
Grupos = c( rep( "Controle", 15 ), rep( "Esteroide", 14 )  )
Tempo = c( 1,2,3,3,3,5,5,16, 16,16,16,16,16,16,16,1,1,1,1,4,5,7,8,10,10,12,16,16,16  )
Censura = c( 0,0,1,1,0,0,0, rep(1, 11), 0,0,1,1,1,1,0,0,1,1,1  )

dados = data.frame( Grupos, Tempo, Censura )

## Selecionando apenas dados sem censura

dados = dados %>% filter( Censura == 1 )
```

Técnicas não paramétricas em sobrevivência

Anteriormente, vimos como estimar a função de sobrevivência e as demais quando não existe a presença de censura. Entretanto, na grande maioria dos casos vamos ter dados censurados, consequentemente teremos que usar técnicas que lidem com dados com censura. Nesta seção, vamos ver técnicas não paramétricas adequadas para esse tipo de situação.

O estimador Kaplan-Meier, também conhecido como estimador produto-limite, será utilizado para estimar a função de sobrevivência, $S(t)$, e o estimador de Nelson-Aalen para estimar a função de risco acumulado, $\Lambda(t)$. Na abordagem não paramétrica não é feita qualquer suposição sobre a distribuição probabilística do tempo de sobrevivência T.

Estimador de Kaplan-Meier

O estimador de Kaplan-Meier é usado para estimar a sobrevivência ($S(t)$) quando existe a presença de censura nos dados. Esse estimador é uma adaptação da função de sobrevivência empírica definida anteriormente quando não temos censuras.

Observação: *Quando não existe censuras o estimador de Kaplan-Meier é igual ao definido na seção anterior.*

O estimador de Kaplan-Meier, na sua construção, leva em consideração as informações censuradas também. Quando analisamos dados com censura, a informação sobre a sobrevivência até a data de sua saída é acessível, isto é, podemos calcular $S(t)$, para todo $t <$ data de saída. Entretanto, quando olhamos para um t maior que a data de saída, não podemos mais garantir se o indivíduo é um sobrevivente ou não, pois não estava mais sendo acompanhado.

O estimador de Kaplan Meier é proposto com o objetivo de acrescentar a informação de censuras na estimação da sobrevivência. Ele utiliza os conceitos de independência de eventos e de probabilidade condicional para caracterizar a curva de sobrevivência em cada intervalo de tempo anterior a t.

Construção do estimador KM

Sejam $t_1 < t_2 < \cdots < t_m$ os m diferentes tempos onde ocorreram os eventos em uma amostra com n indivíduos. Denota-se $R(t_j)$ o número de pessoas no grupo de risco no tempo t_j e $\Delta N(t_j)$ o número total de eventos ocorridos precisamente em t_j. Assim, para os m tempos t_j em que ocorre um evento, a probabilidade de sobrevivência será estimada pelo número de sobreviventes no tempo t_j $(R(t_j) - \Delta N(t_j))$ sobre os que estavam em risco naquele tempo $(R(t_j))$. Como os eventos são independentes, a função de sobrevivência $S(t)$ é estimada empiricamente pelo produto das probabilidade de sobrevivência a cada tempo $t_j \leq t$:

$$\begin{aligned}\widehat{S}_{\mathrm{km}}(t) &= \frac{R(t_1) - \Delta N(t_1)}{R(t_1)} \times \frac{R(t_2) - \Delta N(t_2)}{R(t_2)} \times \cdots \\ &\times \frac{R(t_m) - \Delta N(t_m)}{R(t_m)} = \\ &= \prod_{j:t_j \leq t} \frac{R(t_j) - \Delta N(t_j)}{R(t_j)} = \\ &= \prod_{j:t_j \leq t} 1 - \frac{\Delta N(t_j)}{R(t_j)}.\end{aligned}$$

A mesma equação pode ser apresentada na forma recursiva como:

$$\widehat{S}_{\mathrm{km}}(t_j) = \widehat{S}_{\mathrm{km}}(t_{j-1}) \times \frac{R(t_j) - \Delta N(t_j)}{R(t_j)}$$

PRINCIPAIS PROPRIEDADES DO ESTIMADOR KM:

- É o estimador de máxima verossimilhança de S(t);
- É não-viciado para amostras grandes;
- É fracamente consistente;

- Converge assintoticamente para um processo Gaussiano.

Exemplo: Tempo de sobrevivência (dias) de alguns pacientes da coorte de Aids do Ipec.

$$60\ 84\ 25+\ 54\ 80+\ 37\ 18\ 29\ 50+\ 83\ 80$$
$$81+\ 35\ 52\ 21\ 40\ 22\ 85+\ 39\ 16\ 21+$$

Neste exemplo foram avaliados 21 indivíduos, dos quais 15 morreram (falhou) e 6 foram censurados.

Para calcular a sobrevivência pelo estimador de KM, temos que primeiro obter os tempos de sobrevivência sem censuras, que correspondem ao tempo t_j, com $j = 1, \ldots, 15$. Na tabela seguinte, é possível ver os resultados das quantidades estimados pelo modelo KM.

Tab: Estimacão de sobrevivência por Kaplan-Meier

t_j	$R(t)$	$\Delta N(t)$	$\widehat{S}_{\text{km}}(t) = \prod_{t_j \leq t} \frac{R(t_j)-\Delta N(t_j)}{R(t_j)}$
16	21	1	$\left(\frac{21-1}{21}\right) = 0,9524$
18	20	1	$0,9524 \times \left(\frac{20-1}{20}\right) = 0,9048$
21	19	1	$0,9048 \times \left(\frac{19-1}{19}\right) = 0,8571$
22	17	1	$0,8571 \times \left(\frac{17-1}{17}\right) = 0,8067$
29	15	1	$0,8067 \times \left(\frac{15-1}{15}\right) = 0,7529$
35	14	1	$0,7529 \times \left(\frac{14-1}{14}\right) = 0,6992$
37	13	1	$0,6992 \times \left(\frac{13-1}{13}\right) = 0,6454$
39	12	1	$0,6454 \times \left(\frac{12-1}{12}\right) = 0,5916$
40	11	1	$0,5916 \times \left(\frac{11-1}{11}\right) = 0,5378$
52	9	1	$0,5378 \times \left(\frac{9-1}{9}\right) = 0,4781$
54	8	1	$0,4781 \times \left(\frac{8-1}{8}\right) = 0,4183$
60	7	1	$0,4183 \times \left(\frac{7-1}{7}\right) = 0,3585$
80	6	1	$0,3585 \times \left(\frac{6-1}{6}\right) = 0,2988$
83	3	1	$0,2988 \times \left(\frac{3-1}{3}\right) = 0,1992$
84	2	1	$0,1992 \times \left(\frac{2-1}{2}\right) = 0,0996$

Código R

```
library(survival)
```

```
Warning: package 'survival' was built under R version 4.1.1
```

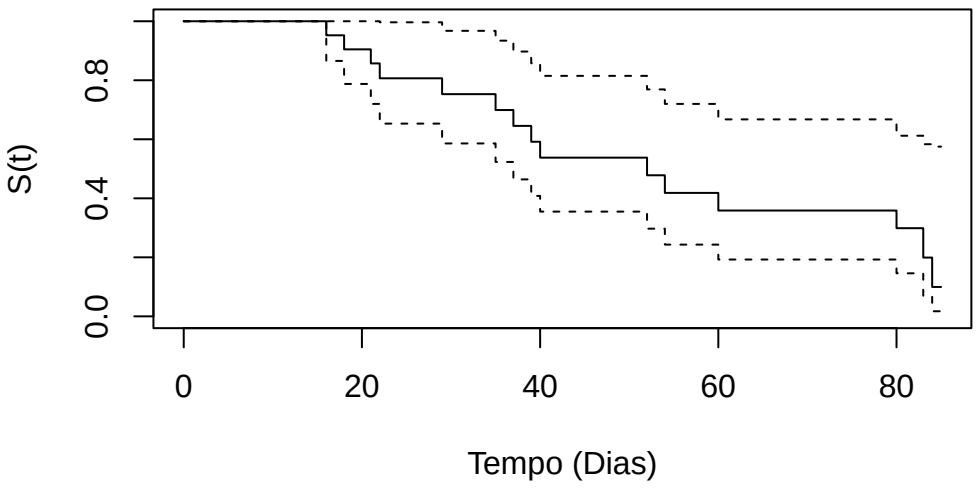
```
tempo = c(60, 84, 25, 54,80,37,18,29,50,83,80,81,35,52,21,40,22,85,39,16,21)
censura = c(1,1,0,1,0,1,1,1,0,1,1,0,1,1,1,1,1,0,1,1,0)
dados = data.frame( tempo, censura )
View(dados)
```

```
mod = survfit( Surv( tempo, censura   )~1, dados  )
summary(mod)
```

```
Call: survfit(formula = Surv(tempo, censura) ~ 1, data = dados)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   16     21       1   0.9524  0.0465        0.8655        1.000
   18     20       1   0.9048  0.0641        0.7875        1.000
   21     19       1   0.8571  0.0764        0.7198        1.000
   22     17       1   0.8067  0.0869        0.6531        0.996
   29     15       1   0.7529  0.0963        0.5859        0.968
   35     14       1   0.6992  0.1034        0.5232        0.934
   37     13       1   0.6454  0.1085        0.4642        0.897
   39     12       1   0.5916  0.1120        0.4082        0.857
   40     11       1   0.5378  0.1140        0.3550        0.815
   52      9       1   0.4781  0.1160        0.2972        0.769
   54      8       1   0.4183  0.1158        0.2431        0.720
   60      7       1   0.3585  0.1137        0.1926        0.667
   80      6       1   0.2988  0.1093        0.1459        0.612
   83      3       1   0.1992  0.1092        0.0680        0.583
   84      2       1   0.0996  0.0891        0.0172        0.575
```

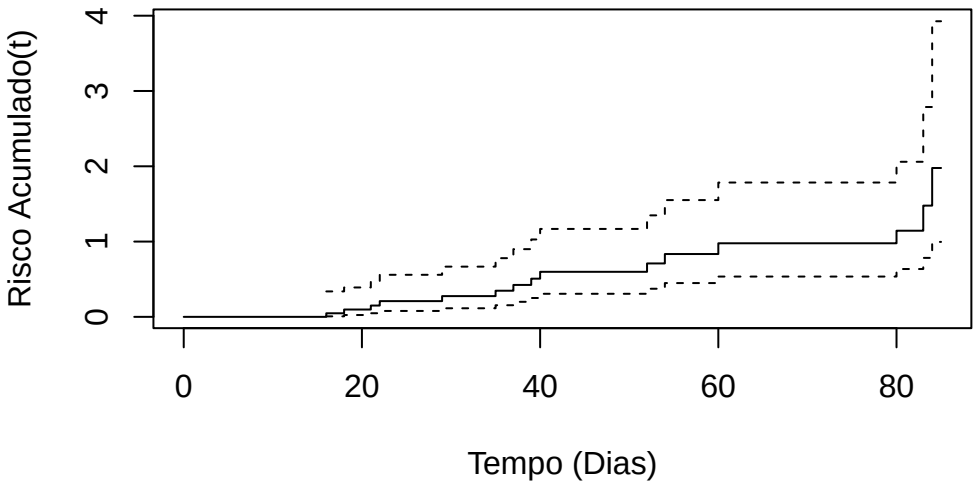
```
plot(mod, xlab = 'Tempo (Dias)', ylab= 'S(t)')
```



A partir da estimação da sobrevivência podemos encontrar o estimador do risco acumulado de Kaplan-Meier:

$$\hat{\Lambda}_{KM}(t) = -\ln\left(\hat{S}_{KM}(t)\right)$$

```
plot(mod, xlab = 'Tempo (Dias)', ylab= 'Risco Acumulado(t)', fun = 'cumhaz')
```



Estimador Nelson Aalen

Uma alternativa para estimar o risco acumulado é o estimador de Nelson-Aalen. Ele foi proposto por Nelson (1972) e retomado por Aalen (1978), no qual demonstrou as propriedades assintóticas do estimador. Esse estimador pode ser obtido através da seguinte expressão:

$$\tilde{\Lambda}_{NA}(t) = \sum_{j:t_j<t} \left(\frac{\Delta N(t_j)}{R(t_j)} \right)$$

Dessa forma, também podemos obter o estimador de $S(t)$ de Nelson-Aalen, tal que,

$$\tilde{S}_{NA}(t) = \exp\{-\tilde{\Lambda}_{NA}(t)\}$$

Comparação de curvas de sobrevivência

O problema de **comparação de distribuições de sobrevivência** surge com frequência em estudos de sobrevivência. Por exemplo, pode ser de interesse comparar dois tratamentos para uma determinada doença.

Um caminho simples é a observação do gráfico das funções de sobrevivência estimadas. Contudo esse gráfico fornece apenas uma idéia aproximada da diferença entre essas distribuições. Ele

não revela se as diferenças são significativas. Para comparar as curvas de sobrevivência mais formalmente, podemos recorrer a **testes de hipóteses.**

Teste Log-Rank

Para confirmar estatisticamente se as curvas de sobrevivência entre grupos são realmente distintas, vamos recorrer aos testes de hipóteses. O teste log-Rank (ou Mantel- Hanzel) é usado para essa finalidade.

Como funciona o teste ?

O teste de log-rank compara a distribuição da ocorrência dos eventos observados em cada estrato com a distribuição que seria esperada se a incidência fosse igual em todos os estratos.

Se a distribuição observada for equivalente à distribuição esperada, dizemos que a curva de sobrevivência dos pacientes pertencentes ao estrato é equivalente à curva de sobrevivência dos pacientes em geral (a covariável não tem efeito na sobrevida).

Realização do teste:

Hipótese nula: não há diferença entre estratos.

Estima-se o número de eventos esperados para cada estrato k, segundo a hipótese nula de incidência igual em cada estrato.

Distribuição esperada de eventos igual em todos os estratos:

$$E_k(t) = \Delta N(t) \frac{R_k(t)}{R(t)}$$

em que $\Delta N(t)$ é o número total de eventos observados. $R_k(t)$ é o número de pessoas em risco no estrato k. $R(t)$ é o número total de pessoas em risco no estudo no tempo t.

Estatística de teste log-rank para dois estratos $(k=2)$:

$$\text{Log-rank } = \frac{(O_1 - E_1)^2}{\text{Var}\,(O_1 - E_1)}$$

O_1 = total de eventos observados no estrato 1

E_1 = total de eventos esperados no estrato 1 .

$$\text{Var}\,(O_1 - E_1) = \sum_t \frac{R_1(t) R_2(t) \Delta N(t) [R(t) - \Delta N(t)]}{R(t)^2 [R(t) - 1]}$$

A estatística log-rank, sob a hipótese nula, segue uma distribuição χ^2, com $k-1$ graus de liberdade.

Observação: Apesar de introduzir o teste log-rank para dois estratos, a sua generalização para k estratos também é válida.

Teste de Peto

Dá maior peso às diferenças (ou semelhanças), no início da curva, onde se concentra a maior parte dos dados e por isso é mais informativa. Usa um ponderador $S(t)$ no estimador.

$$\text{Peto} = \frac{(O_1 - E_1)^2}{\text{Var}(O_1 - E_1)}$$

sendo que

$$O_1 - E_1 = \sum_{t_j} S(t_j)(O_1(t_j) - E_1(t_j))$$

Também a estatística Peto segue aproximadamente uma distribuição χ^2 com $k-1$ graus de liberdade.

A variância da estatística de Peto é igual a variância do log-rank, onde a cada tempo se pondera pelo quadrado da função de sobrevida.

No R

Para fazer comparações de curvas de sobrevivência no R, basta usar a função `survdiff( ).`

Exemplo: Dados de Hepatite

```
tempo<- c(1,2,3,3,3,5,5,16,16,16,16,16,16,16,16,1,1,1,1,4,5,7,8,10,10,12,16,16,16)
cens<-c(0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,1,1,1,1,0,0,0,0,0)
grupos<-c(rep(1,15),rep(2,14))

## Para rodar o teste log-rank, faça rho = 0

survdiff(Surv(tempo,cens)~grupos,rho=0)
```

```
Call:
survdiff(formula = Surv(tempo, cens) ~ grupos, rho = 0)

          N Observed Expected (O-E)^2/E (O-E)^2/V
grupos=1 15        2     4.81      1.64      3.67
grupos=2 14        7     4.19      1.89      3.67

 Chisq= 3.7  on 1 degrees of freedom, p= 0.06
```

```r
## Para rodar o teste Peto, faça rho = 1

survdiff(Surv(tempo,cens)~grupos,rho=1)
```

```
Call:
survdiff(formula = Surv(tempo, cens) ~ grupos, rho = 1)

          N Observed Expected (O-E)^2/E (O-E)^2/V
grupos=1 15     1.79     4.16      1.35      3.43
grupos=2 14     6.00     3.63      1.54      3.43

 Chisq= 3.4  on 1 degrees of freedom, p= 0.06
```